PROPOSTA DE UM SISTEMA DE EDIÇÕES ELETRÔNICAS E SEUS ASPECTOS TECNOLÓGICOS: a construção de um *Piloto de Corpus Eletrônico* a ser implantado na UEFS para o estudo da língua portuguesa no semi-árido baiano (séculos XVII-XXI)

Zenaide de Oliveira Novais Carneiro
1658/2009

Projeto de Pesquisa para estágio de Pós-doutorado 2
2009-2010

INSTITUTO DA LINGUAGEM/Departamento de Lingüística/UNICAMP

Representante Legal da UEFS: Reitor José Carlos Barreto de Santana

Palavras-Chave
1. Processamento tecnológico de linguagem
2. Linguagem XML
3. Ferramenta e-dictor
4. Corpus digital

Área de Conhecimento: Lingüística
Sub-área de conhecimento: Lingüística Histórica

1. Resumo do Projeto:

Este projeto de pós-doutorado, a ser realizado entre 2009-2010, junto ao Instituto de Linguagem da Unicamp, sob a supervisão da Profa Charlotte C. Galves, será desenvolvido no âmbito do "Projeto *Corpus* do Português Histórico Tycho Brahe" e do projeto interdisciplinar "Padrões Rítmicos, Fixação de Parâmetros e Mudança Lingüística - Fase II" (http://www.tycho.iel.unicamp.br/~tycho/). Um projeto que desenvolve metodologias para formação de grandes bancos de dados, segundo os mais recentes avanços do campo do processamento de linguagem natural, em tecnologias próprias da Inteligência Artificial com repercussões em diferentes áreas de conhecimento. O principal objetivo deste estágio é consolidar uma parceria para troca e aperfeiçoamento de novas tecnologias de processamento de texto para construção de um Corpus Digital de um amplo conjunto de dados inéditos de toda grande área do semi-árido baiano, desenvolvido sob minha supervisão ao longo de dez anos de pesquisa na UEFS, tratados sob rigor filológico, através de parceria com a UFBA/IL, no Projeto Para a História do Português (PROHPOR), http:www.prohpor.ufba.br, mas com acesso ainda limitado a meios impressos. Essa parceria com a Unicamp visa a otimização e potencialização desse material, através de aquisição de novas tecnologias, com o uso da linguagem XML e da ferramenta integrada de anotação de corpus, e-dictor (cf http://oncoto.dyndns.org:44880/projects/edictor), desenvolvido por Faria, Kepler e Paixão e Souza (2004-2009).

Justificativa do Orientador para a Concessão da Bolsa:

Tenho grande interesse em orientar esse projeto de programa de pós-doutorado (2009-2010) no Departamento de Lingüística, do Instituto de Linguagem, da Universidade

Estadual de Campinas (UNICAMP), a ser desenvolvido no âmbito do Projeto Corpus do Português Histórico Tycho Brahe e do projeto interdisciplinar Padrões Rítmicos, Fixação de Parâmetros e Mudança Lingüística - Fase II http://www.tycho.iel.unicamp.br/~tycho/ que coordeno. É fundamental que sejam consolidadas parcerias para troca e aperfeiçoamento de novas tecnologias de processamento de texto para construção de corpora digitais eletrônicos nos moldes do que está sendo proposto. Essa abordagem multidisciplinar une o projeto às tendências mais recentes no campo dos estudos da mudança lingüística, aliando o recurso a grandes volumes de dados, contribuindo, por fim, na renovação das perspectivas para o campo. Durante esse período, a candidata poderá se inteirar dos métodos que estamos usando na construção do "Projeto *Corpus* do Português Histórico Tycho Brahe", em particular no que diz respeito à formatação dos textos em linguagem XML e do uso do e-dictor e à anotação sintática dos textos, bem como das ferramentas de busca com que trabalharmos para a recuperação dos enunciados relevantes. Essa ferramenta lhe será certamente de grande valia para o seu projeto de pesquisa na UEFS. Por outro lado, será a ocasião de estreitarmos nossos laços de pesquisa em particular no que diz respeito aos estudos comparativos da diacronia do português, europeu e brasileiro. A concessão de bolsa de pós-doutorado é fundamental para que o projeto se efetive, dada a distância entre as duas instituições. O investimento terá retorno pela relevância e pela possibilidade de se tornar um projeto de grande alcance, com a implantação do Corpus Digital Eletrônico, a partir do plano para que seja executado o projeto piloto na instituição de origem, através da parceria estabelecida com o "Projeto Corpus do Português Histórico Tycho Brahe".

CPF do Orientador:

Orientador: Charlotte Marie Chambelland Galves

Impacto Científico:

O estudo da mudança lingüística, com fortes repercussões na área da linguagem, constitui-se a gênese da própria lingüística como ciência, iniciada em fins do século XIX, e avança com suas metodologias de acordo com o desenvolvimento da própria ciência em geral, tanto do ponto de vista social, biológico cerebral quanto no plano tecnológico na área da inteligência artificial, beneficiados com a automação tecnológica. Como se trata de um campo que cresce com a troca de tecnologias e essas demandam grandes investimentos financeiros e acadêmicos, diversas universidades brasileiras vêm fazendo parcerias em todas as áreas de conhecimento e principalmente na construção de grandes bancos de dados compartilhados mundialmente na rede de computadores, como um recurso imprescindível para fazer lingüística na atualidade. No campo da Lingüística Histórica, no Brasil, um dos pioneiros nesse ramo é o projeto "Corpus Histórico do Português Tycho Brahe" (http://www.tycho.iel.unicamp.br/~tycho), inserido no projeto interdisciplinar "Padrões Rítmicos, Fixação de Parâmetros & Mudança Lingüística", financiado pela FAPESP e CNPq, iniciado em 1998 (cf. relatórios em http://www.tycho.iel.unicamp.br/~tycho/prfpml/fase2/relats.html), na qual me filiei, através da UEFS, desde 2005 (cf. (http://dgp.cnpq.br/buscaoperacional/detalhepesq.jsp?pesq=5992506414152580). No caso desse projeto em particular, vem fazendo parcerias importantes com diversas

instituições internacionais. A colaboração científica nessa parceria que nos propomos é importante porque esse tipo de banco de dados encontra problemas de adequação entre materiais editados na tradição filológica e sua conversão para linguagens computacionais como a XML. No atual estágio, grandes e importantes discussões têm sido feitas com centros de excelência em filologia como a UFBA, no qual faço parte, através do PROHPOR (www.prohpor.ufba.br) com diversas instituições, com o projeto nacional PHPB (conforme relatório http://www.mundoalfal.org/Ataliba%20T.htm). Entrei nessa discussão através do meu projeto de doutoramento na Unicamp, e continuo desde o seu término em 2005. No momento atual, passamos a colaborar com material para alimentação do "Corpus Histórico do Português Tycho Brahe" (ver cartas brasileiras (UEFS) atas de africanos (UFBA), itens 45-48 http://www.tycho.iel.unicamp.br/~tycho/corpus/index.htm. Mas queremos dar um passo a frente e trocarmos tecnologias para implantação e um banco na UEFS em parceria com a UNICAMP. Os ganhos científicos se darão, principalmente, através do aprimoramento de técnicas de conversão para a XML ou Linguagem de Marcação Extensível (*extensible markup language, cf.* W3C, 1997), sem prejuízo do rigor filológico, desenvolvido com excelência na UEFS e na UFBA e agora também em um projeto que se inicia na UESB (com Prof. Visitante Namiutti, proveniente da Unicamp) formando uma parceira importante nesse cenário, unindo tecnologia e tradição.

Impacto Tecnológico:

O avanço científico de modo geral depende crucialmente da manipulação segura e rápida de um grande volume de dados com entrecruzamento com redes espalhados por todo mundo, como uma necessidade global. Diversos projetos na área de lingüística para formação de grandes bancos de dados vêm avançando rapidamente. São exemplos desse o próprio projeto "Corpus Histórico do Português Tycho Brahe", que é filiado ao "Penn-Helsinki Parsed Corpus of Middle English", coordenado por Anthony Kroch, que vem se unindo em uma rede de colaboração mútua via conhecimento de experiências em outros grupos de pesquisa, no Brasil e no exterior, a exemplo do "Comportamento estocástico, Fenômenos Críticos e Identificação de Padrões Rítmicos nas Línguas Naturais (Pronex/Fapesp)", e *o* Projeto "Lácio-Web do Núcleo Interdisciplinar de Lingüística Computacional (NILC), além de contar com o suporte computacional com a rede IME-USP. Sendo assim, o projeto Piloto de Corpus Digital que estamos propondo desenvolver tem como impacto a sua inserção em uma rede internacional de construção de grandes *corpora* históricos anotados. A parceria que proponho através da troca de experiência com o maior centro desse tipo no Brasil é uma necessidade para a Bahia. A fim de que possa contribuir, através da conversão de seus dados armazenados em projetos locais para linguagens mais adequadas. Os avanços se darão em várias frentes. No plano da pós-graduação (mestrado) na UEFS, em fase de implantação, que poderá já sair com grau de excelência na área de grandes bancos de dados, além da iniciação precoce tanto na própria universidade quanto no ensino médio. Também será importante dar visibilidade a materiais já prontos, como será mostrado adiante e que estão sendo divulgados em ritmo lento por dificuldades no manuseio, dinamizando as pesquisas na UEFS, ao tempo que as insere nos contextos nacionais e internacionais. Esse é um projeto que por ser definido na área da tecnologia de informação poderá contribuir em várias áreas de conhecimento, já que a linguagem proposta fornece versões finais diversas que podem ser escolhidas de acordo com o interesse do pesquisador, desde um material conservador, moderno ou o acesso de estruturas etiquetadas (morfológica e sintaticamente) para usos quantificacionais na

lingüística. Isso poderá ser medido através do aumento da eficiência, em números e qualidade de projetos individuais, artigos, etc, dos pesquisadores envolvidos na UEFS.

Impacto Econômico:

Através da discussão até agora feita, viu-se que o tipo de projeto apresentado na área da tecnologia de informação pode vir a contribuir para minimizar os custos futuros. O tipo de banco de dados digital em XML, embora com as dificuldades iniciais de implantação, motivadas principalmente pelas adequações em projetos pioneiros, pode sem dúvida ter um grande impacto econômico no sentido positivo, porque poderá gerar produtos com crescimento exponencial através de redes de cooperação e entrecruzamento de dados com custos cada vez menores. São materiais que podem se tornar disponíveis na rede mundial de computadores, com recuperação rápida e confiável de grandes quantidades de informação com custos inferiores àqueles manuseados por via impressa (livros digitais, e materiais de uso *on line*), etc. São inúmeras as contribuições com retornos cada vez maiores, seja em termos de produtos acadêmicos (teses, artigos, etc), seja através do seu uso mais direto pela comunidade (acessos *on line*), tanto de forma simétrica quanto assimétrica através de ganhos com as facilidades tecnológicas no manuseio das informações. Em resumo, de forma mais iminente, a cooperação entre projetos, como a que estamos propondo, permite reduzir a redução de custos através do compartilhamento de recursos tecnológicos e operacionais.

Impacto Social:

A aquisição de metodologia desenvolvida para a elaboração de grandes *corpora* anotados de língua vem se firmando como um recurso imprescindível para fazer estudos de mudança lingüística em uma das áreas mais importantes da lingüística hoje e que grandes contribuições vêm dando, inclusive na área educacional, através do conhecimento da língua exteriorizada, também, como instrumento legítimo e historicamente construído. E é nesse sentido que pode contribuir como um impacto social importante qual seja, o conhecimento da língua portuguesa no Brasil, definida como português brasileiro. Uma língua que assume papel de liderança em termos de usuários no chamado mundo lusófono. Conhecer a realidade lingüística do Brasil, do qual tem o grande papel a sociolingüística, desde o seu processo inicial de formação como país nos moldes de hoje, requer a manipulação de um grande volume de informações para legitimar todas as variedades lingüísticas dos povos que constituem as bases de sua formação. O Brasil enfrenta ainda hoje grandes problemas na área educacional, sobretudo com dificuldades na interpretação e leitura, o chamado analfabeto funcional. Parte dessa dificuldade advém do desconhecimento do fosso que separa o Brasil no uso de uma língua padrão e suas legítimas variedades vernaculares que formam o português brasileiro, um contexto de diglossia lingüística com todas as implicações que isso possa ocasionar em termos educacionais. A lingüística histórica fornece a base empírica, ao lado de diversas outras áreas da lingüística, um suporte para uma adequada discussão sobre o problema educacional. Contudo, os avanços científicos na área da lingüística requerem uma base de dados eficiente e rápida que acreditamos possa ser dinamizada com projetos dessa natureza. No que diz respeito diretamente ao projeto proposto, poderá permitir a diminuição da concentração regional da pesquisa, permitindo a difusão do conhecimento entre instituições nacionais e internacionais, além de subsidiar outras do interior do Estado.

Impacto Ambiental:

É sabido que o transporte material de informações nos processos anteriores a tecnologia de informação automatizada e digital, demandou um grande volume de matéria prima ambiental, sobretudo com o uso do papel, que necessita para seu produto final do manuseio de agentes químicos, além dos impactos ambientais diversos, despesas de transporte e inúmeros outras interfaces econômicas. O uso do meio digital requer um volume de informações que não seria mais compatível com a capacidade ambiental nos moldes anteriores, isto é, impressos. Os bancos de dados, na forma como está sendo apresentado, trabalha com milhões de palavras com agilidade de informações e com menor dano ambiental, sobretudo na rede de computadores que independe, inclusive, de qualquer tipo de transporte material. Além disso, o uso de computadores, inclusive domésticos, já é uma realidade e a inserção de informações cotidianas já se encontra em curso e é constitutiva do mundo atual. Desse ponto de vista, a preferência por bancos digitais colabora com uma tecnologia mais limpa.

Objetivo Geral:

Este projeto de pesquisa de pós-doutorado tem como objetivo principal desenvolver um "Piloto de um Corpus Digital Eletrônico" nos moldes do "Corpus Histórico do Português Tycho Brahe" (Unicamp) (http://www.tycho.iel.unicamp.br/~tycho), a partir de um *corpus* já constituído ao longo de dez anos de pesquisa, resumido no projeto "Vozes do sertão em dados: história, povos e formação do português brasileiro", subprojeto de um plano central de trabalho que vem sendo desenvolvido na UEFS e como parte do "Programa para a história do português (PROHPOR)", sediado na UFBA, e de acordo com a agenda do "Projeto Nacional Para a História do Português (PHPB)", cabendo a equipe da UEFS, o levantamento, compilação, edição de documentação e estudos lingüísticos da região semi-árida baiana. Os objetivos do "Piloto de um Corpus Digital Eletrônico" são os seguintes:

1. Implementação desse *corpus* em formato de texto computacionalmente manipulável, em linguagem XML, que permita recuperar informações gráficas de cunho filológico dos documentos originais, além disso, gerar bases anotadas (morfológica e sintática) para análise lingüística. Ou seja, um *corpus* que permita usos de mecanismos que gerem versões diversas acessíveis em processamentos de buscas automáticas;

2. Após o desenvolvimento de um modelo de desse Piloto, o objetivo é implantar o banco com dados do Projeto Vozes do Sertão em Dados, em versão eletrônica manipulável, para uso na pós-graduação e também ser disponibilizado na rede mundial de computadores;

3. Como sub-objetivo ligado ao objetivo 2, gerar léxicos das edições em XML, catálogos dos documentos, edições variadas (conservadoras, modernizadas, etc), além de outros subprodutos, todos gerados pela transformação XML.

Objetivos Específicos:

Os objetivos específicos estão ligados as metas e aos posteriores resultados desse projeto, a saber:

1. Acompanhar o início da alimentação do "Corpus Histórico do Português Tycho Brahe" (http://www.tycho.iel.unicamp.br/~tycho), no que se refere ao material editado para análise do volume 2, da minha tese de doutorado intitulada *Cartas brasileiras (1809-1904): um estudo lingüístico-filológico* de forma a que venha servir de material para ser pensado um *corpus* piloto (cf. itens 45-48 http://www.tycho.iel.unicamp.br/~tycho/corpus/index.htm);
2. Estabelecimento de um *corpus* inicial formado por cartas escritas por brasileiros cultos nascidos entre fins do século XVIII e brasileiros cultos nascidos até o terceiro quartel do século XIX a ser editada em XML;
3. Codificação XML inicial de 185 cartas a partir de um *corpus* editado em Word segundo de acordo com as normas da tradição filológica para uso lingüístico histórico;
4. Uso do e-dictor para controle do texto base bruto e transformação XML;
5. Testes da codificação XML;
6. Anotação morfológica através da equipe técnica da Unicamp do Projeto "Corpus Histórico do Português Tycho Brahe"Estabelecimento do formato da base de dados e do sítio-web do projeto piloto final;
7. Teste do projeto Piloto de forma que venha ser possível a alimentação de textos até atingir milhões de palavras em projeto contínuo na UEFS com metodologia do projeto base gerido pelo "Corpus Histórico do Português Tycho Brahe"/Unicamp.

Metodologia:

Como se trata de um plano de conversão de um *corpus* já parcialmente construído, faz se necessário demonstrar a metodologia já aplicada no banco bruto e depois demonstrar o plano propriamente dito em um Corpus Digital Eletrônico, pensado inicialmente a partir de um piloto a ser executado como está sendo posposto nesse estágio de pós-doutoramento na UNICAMP. Dessa forma, parto inicialmente para a apresentação do estágio atual do banco de dados que coordeno na UEFS com o objetivo de mostrar em que medida é necessária a transformação para um Corpus Digital Eletrônico e como isso será executado passo a passo. Nesse ponto, será necessário fazer um balanço breve do material das pesquisas feitas na UEFS em parceria com a UFBA, através do "Projeto PROHPOR, Programa Para a História do Português", e também com uso das metodologias desenvolvidas nos últimos dez anos do Projeto Nacional, integrado com várias universidades, o PHPB (Para a História do Português Brasileiro), a saber: UFPE, Bahia UFBa/UEFS, UFPb, UFAL, UFCE, UFMG, UFRJ (http://www.letras.ufrj.br/phpb-rj/,) UEL, UFSC, USP e UNICAMP.

1. DA PARTE DO CORPUS COMPILADO PELA UEFS (1998-2008) e de outras etapas previstas:

O semi-árido é uma região ainda pouco estudada do ponto de vista lingüístico e histórico. Neste item, apresentamos o resultado de dez anos de pesquisa, referente a estudos sócio-históricos e compilação de documentação na região de forma a permitir estudos na área da lingüística histórica. É um banco que assume uma feição multidisciplinar, envolvendo pesquisas históricas e lingüísticas, além, é claro, da interface com a filologia, que fornece todo o aparato para o tratamento da documentação. Da parte histórica, é fundamental o que chamamos de controle de produção do material, ou seja, todas as classificações que, necessariamente, precisam ser feitas com relação ao tipo de material a ser analisado. De modo geral, para além das contribuições nas áreas já citadas, oferece uma contrapartida às regiões estudadas, por colaborar na preservação da cultura regional e oferecer subsídios para uma história lingüística local com implicações sócio-educacionais.

O objetivo é geral desse banco é compor uma amostragem de documentos históricos (século XVII-XX), editados para fins lingüísticos de várias regiões do semi-árido baiano, a saber: Nordeste, Serra Geral, Chapada Diamantina e Paraguaçu, etc., para entender como se deu o avanço e a consolidação da língua portuguesa nessa região. São regiões que exemplificam o processo de ocupação européia falante do português na região. Através da extensão e consolidação de uma fase já concluída que trata de formação de banco de dados e análise lingüística de natureza sócio-histórica.

O segundo objetivo é usar o *Corpus* para estudar a história do português a partir de questões levantadas por projetos em Lingüística histórica no que concerne a história do português europeu, como o Programa para a História do Português/PROHPOR. As questões para o português brasileiro estão sendo elaboradas com base no conhecimento acumulado, em dez anos de pesquisa, pelo PHPB, do qual fazemos parte. Essas questões são as seguintes: 1. quais são as características da gramática do português europeu **no** Brasil? 2. qual é a trajetória no tempo dessa gramática? 3. como se dá a emergência **do português brasileiro**? Essas questões já podem ser formuladas, uma vez que foi composto um vasto material de pesquisa inédito e fidedigno com análises já feitas na documentação do século XIX. Dessa forma, o problema da mudança nos textos escritos por brasileiros, em diferentes períodos do português, torna-se mais apreensível, uma vez que o projeto já detém um banco de dados sólido. O terceiro objetivo é propiciar um amplo conhecimento sócio-cultural e histórico de povos que viveram na região do semi-árido baiano, que embora seja conhecida por seu aspecto homogêneo climático, abriga uma rica diversidade geo-ambiental com impactos importantes no processo de ocupação e domínio de população de origem portuguesa e outras em detrimento de população autóctone durante todo o período colonial. Esse banco poderá fornecer meios para o resgate, não apenas documental, como o descrito, a seguir, mas também da diversidade lingüística da região. O quarto objetivo é o de resgatar documentação dispersa no Brasil e no exterior, gerada no semi-árido, uma vez que pouco ou quase nada se encontra em comunidades locais, fato que em muito contribuirá para a preservação da história regional, pouco documentada.

A metodologia seguiu as seguintes fases:

(i) seleção e identificação da autenticidade dos documentos; localização espacial e temporal cuja importância, do ponto de vista documental, é fundamental porque busca a preservação e disponibilidade de material para análise em diversas áreas de conhecimento, além da lingüística; (ii) identificação de dados relevantes

sobre os escreventes e destinatários; (iii) contextualização da amostra com base na história externa do português brasileiro e (iv) edição em versão conservadora para preservar marcas lingüísticas relevantes para o estudo histórico do português. O produto final é apresenta-se a seguir separados por fases: concluída, em andamento e previstas:

I. CONCLUÍDOS

1. 2008-2009. "Análise lingüística em documentos da imprensa feirense, 1900-2000" –Imprensa: 120 cartas de leitores e redatores publicadas na imprensa feirense entre 1900-2000.

2. 2008-2008 "Cartas pessoais de cultos e semi-cultos escritas na primeira metade do século XX: estudos lingüísticos e sócio-históricos". - 230 cartas pessoais do século XX .

3. 2000-2005 "Cartas brasileiras (1808-1904): um estudo lingüístico filológico. Res.- 500 cartas particulares do séc. XIX

4. 2000-2000 "Cartas de leitores na imprensa baiana do século XIX: composição de corpus e análise lingüística" - 19 cartas de leitores publicadas em jornais feirenses.

5. 1999-1999 "Anúncios na imprensa baiana: um corpus de pesquisa sobre o português escrito em jornais do século XIX" -: 320 anúncios, sendo 100 os publicados em jornais feirenses

6. 1997-1999 "Contribuição para a constituição de um banco de textos e de um banco de dados para o estudo da história da língua portuguesa, no Brasil, do século XVII-XX" - 30 documentos (inventários, testamentos e declarações - século XVII e século XX)**.**

II. EM ANDAMENTO

1. 2008-2010 "Compilação de possíveis fontes escritas, em português, por "Tapuia" no interior da Bahia (séc. XVII-XVIII)"

2. 2008-2009. "Resgate dos documentos do sobrado do Brejo Seco, Serra Geral – documentos pessoais/ Resgate de documentos de Campinas (Dr. Lycurgo e Recife, da Fundação Gilberto Freire.

III - PROJETOS PREVISTOS

1. 2010-2010 – "Resgate dos documentos do sobrado do Brejo Seco, Serra Geral - Fase II: **Livros do "Campo Seco":** *anotações de* três gerações sertanejas baianas (1755-1832)" -

2. 2011 – 2011 - Cartas pessoais de cultos e semi-cultos escritas feirenses, séculos XX: estudos lingüísticos e sócio-históricos" -

3. 2010-2010 - "Documentação de irmandades de pretos e pardos do semi-árido baiano"

2 DA PARTE DO PILOTO DO CORPUS DIGITAL A SER FEITO NA UNICAMP PARA POSTERIOR IMPLANTAÇÃO NA UEFS

A metodologia do Projeto *Piloto de Corpus Digital* baseia-se fundamentalmente na metodologia do "Corpus Histórico do Português Tycho Brahe", composto por um corpus eletrônico anotado de textos em português escritos por autores nascidos entre 1435 e 1845, desenvolvido desde 1998. A evolução da maturidade metodológica desse projeto pode ser vista no site: http://www.tycho.iel.unicamp.br/~tycho/corpus/, onde estão definidos as ferramentas e modelos seguidos e que apresentarei de forma resumida. O resultado da aplicação da metodologia já resulta em um produto atual de 52 textos (**2.357.626** palavras), disponíveis para pesquisa livre, com um sistema de anotação lingüística em duas etapas: **anotação morfológica** (aplicada em 25 textos); e **anotação sintática** (aplicada em 5 textos). Atualmente, o projeto vem aplicando a sua metodologia em textos produzidos por brasileiros. A experiência pioneira realizada na minha tese de doutoramento com textos identificados por data de nascimento de autores serviu como um teste para alimentação desse banco com textos escritos por brasileiros, além de outros, como pode ser visualizada na página citada acima, item 48. A demonstração da aplicação sobre conversão de edição em Word em linguagem XML, através da ferramenta e-dictor (cf http://oncoto.dyndns.org:44880/projects/edictor), desenvolvido por Faria, Kepler e Paixão e Souza (2004-2009), demonstrada na "I Oficina do Projeto na UFBA", em 2006, foi o que motivou esse plano de projeto piloto na Unicamp a ser implantado posteriormente na UEFS. A demonstração inicial feita sob a coordenação de Paixão de Sousa pode ser visualizada no seguinte endereço (http://www.ime.usp.br/~tycho/participants/psousa), onde a autora apresenta um primeiro manual usando exemplos de documentos editados em minha tese de doutoramento.

Então, em que consiste o processamento eletrônico de textos? Em seu projeto de pós-doutoramento na Unicamp, Paixão de Souza apresenta um plano para elaboração de *corpus* em linguagem XML. Nesse projeto que intitula de memórias de textos (disponível em http://www.ime.usp.br/~tycho/participants/psousa/memorias/critical_hyper/ece_Frames et.html), a autora, apropriadamente mostra as diferenças nos tipos de edição, opondo o texto manuscrito *versus* do texto editado manualmente em word como se configura o corpus editado no projeto apresentado acima no item (1) e do texto digital/eletrônico *versus* texto produzido eletronicamente. A autora defende apropriadamente que "isso pode ser fundamental no trabalho de natureza filológica - no qual o texto é o objeto central do estudo. Ultimamente, a transcrição e edição de textos antigos têm sido realizadas em meio eletrônico - ao menos, no plano do arquivamento da documentação. Neste trabalho de transcrição e edição, codificam-se as informações relevantes sobre os textos. E mostra como a linguagem **XML** apresenta todas as vantagens da linguagem **HTML, mas com a vantagem fundamental, a de s**ua natureza inteiramente flexível, permitindo o controle e a geração de várias camadas de textos a ser produzida de acordo com a necessidade do pesquisador. Esta anotação na "Linguagem de Marcação Extensível" (*extensible markup language, cf.* W3C, 1997. Na linguagem XML é gerada uma estrutura de dados na forma de árvore, cujos nós são os elementos XML (cf. demonstração partir da qual se pode acessar facilmente a hierarquia das anotações e preencher seus valores nos campos adequados da interface de janelas da ferramenta. Para demonstração é possível acessar

http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/prep/index.html), devido a impossibilidade de inseri-lo aqui. As diretrizes na anotação dos documentos são as seguintes, de acordo Namiuti e Santos (2009): a catalogação dos textos; a transcrição dos textos; a codificação da interferência editorial sobre os textos; (iv) a apresentação dos textos. Por fim, Galves (2004) resume bem as vantagens desse sistema para um "Corpus Digital Eletrônico": a) um **melhor gerenciamento** de arquivos do *corpus*, prevendo sua ampliação no futuro próximo; b) a **otimização dos processos** que levam às anotações morfológica e sintática; c) a **ampliação da finalidade** do corpus, explorando as potencialidades dos textos ortograficamente transcritos e d) 4. a **padronizacão** do *corpus* de modo a poder inserí-lo em catálogos internacionais.

Para facilitar a transformação em XML será usado a Ferramenta Integrada de anotação de corpus, e-dictor, pensada por Paixão de Souza e Fábio Kepler, em 2007, com é feito na Unicamp. Em que consiste essa ferramenta? O **e-dictor** destina-se a transcrição e codificação de textos em formato XML para análises lingüísticas (morfológica, sintática, entre outras). Essa ferramenta, atualmente está sendo implementada por Fábio Kepler (USP) e Pablo Faria (http://oncoto.dyndns.org:44880/projects/edictor) na linguagem de programação Python, com código-fonte aberto a ser disponibilizado à comunidade. O objetivo dos autores, ao adotar o padrão XML, deveu-se a necessidade de se abarcar informações de edição, de etiquetagem (morfológica), além de parte1 do *layout* do texto original (títulos, quebras de linha, página, etc.). Essa ferramenta permite entrada de texto puro a ser convertido para uma estrutura XML; manipulação das principais informações XML: páginas, parágrafos, sentenças e palavras; manipulação das principais edições: opções <ed>, forma fonológica e morfológica; navegação por páginas; layout básico (para exibição das informações na tela); atalhos de teclado para tornar a edição mais eficiente; inserção de comentários de edição; manipulação de metadados; acesso aos demais elementos do texto (títulos, cabeçalho/rodapé, etc.); mais poder de formatação (quebra de seções, títulos e subtítulos, etc.); melhoras gerais no layout da tela da ferramenta; identificação visual dos limites dos elementos, como parágrafos, seções, etc. mais opções na barra de ferramentas, melhor acesso a informação contextual, durante a edição (propriedades de elementos, propriedades do documento, etc.) e melhorar a apresentação HTML do texto.

1. DA PREPARAÇÃO DOS TEXTOS - Tipos de Texto-Fonte e Tipos de Edição. Parte-se do texto-Fonte com grafia preservada (originais impressos e transcrições diplomáticas): e*dição Completa* que implica modernização controlada do texto fonte. Texto-Fonte com grafia modernizada (edições intermediárias, usadas na Fase I): *Edição Técnica* devido as dificuldades de processamento; e - 2. Versões Disponíveis - Versão transcrição do texto-fonte onde mostra a transcrição fidedigna em relação ao texto tomado como fonte (seguindo, portanto sua grafia: a grafia original dos originais impressos; ou a grafia modernizada pelo editor anterior). Versão Texto Editado em que mostra o texto com as interferências realizadas pela equipe do corpus (modernizações completas ou modificações técnicas, conforme o caso). Dois tipos de arquivos estão disponíveis nesses casos: arquivos .html para leitura e arquivos .txt sem formatação, para uso das ferramentas automáticas e busca de dados e, por fim, a versão glossário de edições que mostra uma lista das intervenções realizadas pela equipe do corpus, seja no caso de edições completas ou técnicas. Para correção dos textos há ferramentas disponíveis criada na Universidade da Pensilvânia por Beth

Randall, também autora da ferramenta de busca *CorpusSearch*. *CorpusDraw* e *CorpusSearch* são distribuídos no mesmo pacote no endereço corpussearch.sourceforge.net. *CorpusDraw* (cf. relatório, 2007 http://www.tycho.iel.unicamp.br/~tycho/prfpml/fase2/relatorios/2007/inicio.pdf). O manual para uso do sistema de anotação morfologia e sintática pode ser acessada em http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/index.html.

2. DA ANOTAÇÃO SINTÁTICA E MORFOLÓGICA

Essas ferramentas de anotação automática vêm passando por contínuos aperfeiçoamentos. Inicialmente, Marcelo Finger (IME-USP) desenvolveu um etiquetador automático para o português baseado em parte no etiquetador de Eric Brill para o inglês (cf. Finger 1998, 2000) com precisão de 97% para os textos modernizados.

A elaboração de um analisador sintático segue o sistema de anotação desenvolvido para o inglês no projeto de Anthony Kroch na Universidade da Pensilvânia (cf. http://www.ling.upenn.edu/mideng) no manual de anotação sintática disponível em http://www.ime.usp.br/~tycho/corpus/manual, elaborado inicialmente por Helena Brito, responsável pela fase de treinamento do analisador sintático (*parser*) multilingue desenvolvido por Dan Bikel na Universidade da Pensilvânia, usado pelo Corpus Tycho Brahe, cedido por Anthony Kroch. Como descreve Galves em http://www.tycho.iel.unicamp.br/~tycho/prfpml/fase2/projeto_completo.html, a "seqüência treinamento/correção é própria da metodologia dos analisadores probabilísticos, que segundo Galves parecem mais adequados que os analisadores baseados em regras (como o de Eckard Bick), por serem mais flexíveis e evolutivos. Passemos agora a falar do produto final.

3. DO ACESSO GRATUITO AOS TEXTOS AOS USUÁRIOSATRAVÉS DE CADASTRO

Através de catálogos de listas ordenadas; diretório de textos e consulta aos textos com CorpusSearch, ferramenta criada na Universidade da Pensilvânia por Beth Randall (corpussearch.sourceforge.net). Há, ainda, a possibilidade de se baixar o compactador em ZIP, disponibilizando-se os seguintes produtos Corpus completo, *anotação sintática; corpus completo, anotação morfológica e corpus completo, sem anotação*

O nosso Projeto Piloto pretende passar por todas as fases, incluindo a anotação morfológica e sintática, uma das fases mais importantes para análise lingüística, uma vez que o seu produto é extraído de forma rápida em sistemas de buscas, permitindo que se trabalhar de forma segura com um grande volume de dados, o que dificilmente se conseguiria manualmente, diminuindo incomensuravelmente o tempo de análise nos moldes tradicionais. Contudo essa fase de anotação sintática é a parte mais complexa do nosso plano e depende crucialmente de um aprendizado na Unicamp. No momento, já fizemos uma visita a essa instituição em março de 2009, com recursos próprios, dando início ao nosso projeto de pós-doutoramento. Entretanto, devido os custos serem altos, esse plano vem se tornando inviável, tornando-se crucial o auxílio através de bolsas de pós-doutoramento como a que estamos pleiteando.

Justificativa:

Depois de mais de dez anos de trabalho com a formação de um extenso banco de dados históricos do sertão baiano, com documentos editados por mim ou sob a minha supervisão na UEFS, e como disse no resumo, tratados sob rigor filológico, através de

parceria com a UFBA/IL, no PROHPOR (http:www.prohpor.ufba.br), chegamos a um ponto importante relacionado ao tipo de armazenamento, que é ainda bastante limitado porque apesar de muitos terem sido publicados no âmbito do "Projeto Nacional Para a História do Português Brasileiro. Intentamos possibilitar um uso mais adequado do material, ou seja, o de reformatar o *corpus* diacrônico que já preparamos para que adquira múltiplas finalidades e seja padronizado segundo diretrizes internacionais para a codificação lingüística, em linguagem XML, como propôs Paixão de Souza (2006) em seu projeto de pós-doutoramento na Unicamp, quando promoveu a ampliação e consolidação do *Corpus Histórico do Português Anotado Tycho Brahe* com o uso da tecnologia de processamento dessa linguagem, o que vem possibilidade o controle e a flexibilidade no tratamento computacional a esse corpus, garantindo, ainda, o rigor da descrição lingüística e a multiplicação de seu alcance para diferentes usuários finais. A exemplo do que foi proposto pela autora, também pode ser aplicado ao corpus do sertão, desenvolvido na UEFS, isto é, fazer um piloto para ser aplicado futuramente a todo o corpus que existe além de sua ampliação para dados do século XVII, fase na qual estamos trabalhando, com a coleta de materiais raríssimos, produzidos nos aldeamentos por indígenas. Aqui abro um parêntese para justificar que durante a minha estada na Unicamp, também participarei das discussões com o importante projeto de historiografia indígena coordenada por J. Monteiro, a convite seu (http://www.ifch.unicamp.br/ihb/), já que é um material que está sendo coletado sob a minha supervisão. Então como se vê esse piloto de edição eletrônica será alargado, pois além dos materiais já editados, totalizando mais de 2000 documentos editados ou em fase de finalização, está aberto a novos materiais. O que precisamos é do acesso a novas tecnologias para essa nova fase para dar um tratamento computacional aos textos editados, visando a multiplicação de suas finalidades, tanto do ponto de vista da concepção de um *corpus* histórico para uso lingüístico e educacional, como para diversas áreas de conhecimento, ampliando consideravelmente o uso do material, a exemplo do que vem sendo feito por filólogos como Rita Marquilhas em Portugal (www.clul.ul.pt/equipa/rita_**marquilhas**.php). O sucesso da manipulação dos documentos em linguagem XML já foi amplamente demonstrada como comprovam os relatórios de pesquisa de Paixão de Souza, responsável por sua implantação de (2004-2006)
(http://www.ime.usp.br/~tycho/participants/psousa/memorias/relat_2005/html_files/relat _2005.html) aprovado pela FAPESP, junto ao *Corpus Histórico do Português Anotado Tycho Brahe,* ao qual estamos apresentando o projeto de Piloto do Corpus Digital Eletrônico. Contudo, esse processo de transformação para um requer uma tecnologia que, felizmente já está sendo desenvolvida no Brasil pelo projeto Tycho Brahe, em parcerias com outras instituições internacionais, a exemplo da Penn-Helsinki Parsed Corpus of Middle English/ http://www.ling.upenn.edu/hist-corpora/ na Universidade da Pensilvânia/*USA*. Acreditamos que essa reestruturação que pretendemos implantar no nosso banco de dados na UEFS trará vantagens tanto de otimização dos processos de análise linguística, quanto pela ampliação do escopo de utilização do *Corpus*, como defende Paixão de Souza (2006), uma vez que a linguagem que pretendemos usar no nosso corpus, a *Extended Markup Language* (XML), é a mais usada nos principais *corpora* de língua, para codificar as informações estruturais nos níveis lingüísticos. No plano internacional, os *corpora* eletrônicos são uma das áreas de maior crescimento nos estudos da linguagem e da inteligência artificial. Isso decorre do fato de permitir ao mesmo tempo, acesso a grandes volumes de dados, para diversos fins, além de permitir o desenvolvimento de tecnologias para processamento de linguagens tanto natural quanto artificiais. Esperamos que a UEFS juntamente com a UFBA venham a se tornar

uma referência em *corpora* eletrônicos no nordeste e também que com isso contribua para preservação e difusão de conhecimento científico, histórico e cultural do semi-árido baiano. E que possamos posteriormente fornecer dados para elaboração de léxicos e dicionários, e, como disse, se for possível, encontrar os documentos perdidos escritos pelos indígenas de origem macro-Jê que habitaram e ainda habitam os sertões baianos, buscando pistas de suas línguas quase extintas e de sua influência na formação do português regional baiano.

Justificativa do Coordenador de Curso de Pós-Graduação da instituição de vínculo do candidato para sua participação:

A iniciativa em projetos de mútua cooperação com entre projeto da UEFS, como esse que aqui se apresenta, e o "Projeto Corpus do Português Tycho Brahe" é importante para fortalecer grupos de pesquisas em universidades que ainda se encontram em fases iniciais, sobretudo no que se refere ao desenvolvimento de tecnologias novas e que necessitam de grandes investimentos, como já me refeir. O projeto de capacitação permanente de doutores através de estágios de pós-doutoramento é importante para programas de pós-graduação na Área da Lingüística. Assim, esperamos contribuir com a articulação de uma rede de pesquisa que agregue grupos e instituições e centros tecnológicos.

Resultados esperados:

O Piloto de Corpus Digital Eletrônico tem como principal resultado a aplicação na aplicação de tecnologias que possibilitem a convergência de edições tradicionais e limitadas a usos impressos ou mesmos digitais estáticos, em camadas de textos manipuláveis via uso da linguagem XML e da ferramenta integrada de anotação de *corpus*, e-dictor (cf http://oncoto.dyndns.org:44880/projects/edictor), desenvolvido por Faria, Kepler e Paixão e Souza (2004-2009) sobre o produto armazenado em diversos anos de pesquisa na UEFS, através da metodologia subsidiada, como várias vezes enfatizada nesse projeto, pela técnica desenvolvida pelo projeto "*Corpus Histórico do Português Tycho Brahe"*com o *sistema de edições eletrônicas.* De forma resumida, os resultados serão os seguintes:

1. Iniciar um plano de Corpus Digital Eletrônico permanente na UEFS, a partir da metodologia de edições eletrônicas em XML dos "Corpus Histórico do Português Tycho Brahe**"**; **CNPq - Edital Universal 2007** (Dez/2007 - Dez/2009).
2. Colaborar na metodologia de aplicação do e-dictor (cf http://oncoto.dyndns.org:44880/projects/edictor) para troca mútua de experiências
3. Disponibilização e transferência das metodologias para outros bancos regionais;
4. Subsidiar o *Corpus Digital Eletrônico* de forma que acompanhe os avanços científicos através de avaliações com outros grupos;
5. Disponibilizar formas de acesso a produção automática propiciadas pela transformação XML, tais como léxico das edições, catálogos, etc.
6. Produzir materiais através do uso direto do Corpus Digital Eletrônico na rede mundial de computadores para levantamento extensivo de dados de forma confiável e rápida para uso acadêmico.

7. Elevação do número de artigos dos membros participantes;
8. Subsidiar o programa de pós-graduação *stricto sensu* em Lingüística em fase de implantação na UEFS.

Limitações do projeto:

A principal dificuldade já está sendo vencida que é a preservação da qualidade filológica dos documentos na transformação para a linguagem XML, uma experiência que vem sendo desenvolvida pelo projeto "Corpus Histórico do Português Tycho Brahe/Unicamp" desde 1998. Inicialmente, houve uma resistência no uso de ferramentas computacionais avançadas no trabalho de edição filológica. As discussões vêm progredindo e dirimindo os problemas nas adequações entre a tradição filológica e novas linguagens que, embora desconhecidas entre os lingüistas brasileiros, já que vêm se mostrando capazes de manter a fidelidade do texto sem prejuízo na análise lingüística. Apesar do contato com a Unicamp, inclusive feito na UFBA com a participação de pesquisadores da UEFS, com a primeira da I Oficina de Anotação, em conjunto com o grupo de pesquisas da Unicamp, na UFBa, em abril de 2006 (cf. *PAIXÃO DE SOUSA*, M. C. Manuais de Edição Eletrônica - *Oficinas de Anotação*, Projeto Corpora, 2006) falta um aprendizado sistemático de edições eletrônicas, com ferramentas que exigem treinamento em laboratório. Esse trabalho requer o aprendizado na manipulação linguagens que são novas para lingüistas históricos.

Mecanismos de Transferência de Resultados:

Além da mútua cooperação entre instituições e de partilha de conhecimentos e tecnologias, há outros aspectos:

1) Acesso do banco para execução de trabalhos científicos pela pós-graduação em Lingüística da UEFS;

2) Parceria com a UNICAMP para transferência mútua de tecnologia;

3) Parceria com outros bancos de dados formados por *Corpus Digital Eletrônico*, tanto no Brasil quanto no exterior, para comparação de vertentes do português em diferentes momentos históricos e com outras línguas em geral.

Mecanismos Gerenciais de Execução Multi-Institucional:

O projeto Piloto de Corpus Digital Eletrônico depois de pronto será coordenado pela UEFS, mas com decisões metodológicas e gerenciais tomadas em conjunto com o Projeto "Corpus do Português Histórico Tycho Brahe*"* e do projeto interdisciplinar "Padrões Rítmicos, Fixação de Parâmetros e Mudança Lingüística - Fase II", através de atividades periódicas atualizadas em workshops e oficinas oferecidas com o apoio acadêmico do projeto Tycho Brahe, em uma ou outra instituição, quando necessário. Ou seja, apesar de não requerer um acordo formal entre gestores, esse projeto Piloto de Corpus Digital Eletrônico, desenvolvido no projeto de pós-doutoramento tem metas de implantação na UEFS com mútua cooperação com o Projeto *Tycho Brahe* que pode contribuir com o fortalecimento do grupo de pesquisa na UEFS ao disponibilizar o acesso a novas tecnologias, como já vem sendo feito, embora de forma assistemática. Essa parceria de certa forma já foi firmada através da co-participação no Diretório de Pesquisa do CNPq.

Infra-estrutura disponível:

O Projeto "*Corpus do Português Histórico Tycho Brahe"* e do projeto interdisciplinar "Padrões Rítmicos, Fixação de Parâmetros e Mudança Lingüística - Fase II" é coordenador por Charlotte Marie Chambelland Galves e têm como linhas centrais a análise e modelamento dos sons das línguas naturais; a história da língua portuguesa, lingüística de *corpus*, a modelagem estocástica dos padrões rítmicos das línguas naturais e a sintaxe gerativa das línguas naturais. Define-se, conforme plataforma Lattes, pela e elaboração do *corpus* anotado do português histórico contendo mais de 2 milhões de palavras de autores portugueses e brasileiros nascidos entre os século XV-XIX, disponível em formato eletrônico no site do projeto (http://www.tycho.iel.unicamp.br/~tycho). Apresenta uma equipe formada por 14 pesquisadores, entre os quais me incluo, além de 24 estudantes de Doutorado, Mestrado e Iniciação Científica e o programador de computação e a técnica Cíntia Pano. Em 2003, por exemplo, através de um convênio CAPES/DAAD, com o Departamento de Lingüística da Universidade de Bielefeld, dois especialistas alemães dessa linguagem visitaram o projeto, Jan-Torsten Milde, e Thorsten Trippel. Na segunda visita, Maria Clara Paixão de Sousa, trabalhou nas questões levantadas pela formatação do Corpus nessa linguagem.)". Para a parte de suporte tecnológico computacional conta com a rede IME-USP. Fora do Brasil, filia-se ao pioneiro *Penn-Helsinki Parsed Corpus of Middle English* , coordenado por Anthony Kroch. Apresenta uma completa infra-estrutura na UNICAMP com uma sala específica e com todo o mobiliário no IEL/UNICAMP. A estrutura computacional atual do projeto tem por finalidade atender a dois requisitos principais: suportar a hospedagem do sítio virtual e do corpus de textos, em si; e hospedar aplicações que permitam aos bolsistas manterem o corpus, bem como manipulá-lo para consultas, etc. e apresenta a seguinte configuração:

- 1 servidor Dell, com capacidade para ficar 24hrs no ar, hospedando o sítio virtual do projeto e do corpus, os arquivos do corpus e as ferramentas ligadas a ele; Possui como sistema operacional o Ubuntu (Linux) e é onde os bolsistas do projeto são cadastrados. 2 estações (desktops) de trabalho extras, para bolsistas, com uma delas (a mais moderna) servindo também como espelho do servidor (uma salvaguarda para situações de falha no servidor). 4 laptops (1 permanente, na sala, e 3 para participantes do projeto); Tantos as estações, quanto os laptops, possuem sistema operacional Ubuntu, com todas as ferramentas necessárias instaladas e constantemente atualizadas. O servidor e as estações estão na rede do IEL. Para laptops, há um roteador na sala (ligado à rede do IEL) que fornece acesso Wi-Fi. O no-break específico para o servidor, evitando problemas graves quando há falta de energia.

Revisão de Literatura (Fundamentação Teórica):

Essa parte teórica será organizada em quatro planos para facilitar a apresentação da questão tratada, nos seguintes termos:

1. Da edição de documentos nos moldes filológicos tradicionais; 2. Das questões teóricas do conceito de mudança lingüística; 3. Da análise de fenômenos lingüísticos e 4. Da formação de bancos digitais eletrônicos.

1. Da edição de documentos nos moldes filológicos tradicionais

A questão mais complexa envolvida na construção de grandes *corpora* foi redimensionada em 1997 com a agenda do "**Projeto Para a História do Português Brasileiro"/PHPB/BA/Feira de Santana**, consolidada por ocasião do VII Seminário Nacional do PHPB (Londrina, maio de 2007), tendo sido publicados vários volumes sobre os trabalhos apresentados que culminarão em um conjunto de livros conclusivos sobre intitulados *História do Português Brasileiro* (planejado para 2011-2012), no qual escreveremos dois capítulos. O planejamento e a redação desse livro refletem as agendas das equipes regionais, como por exemplo, a nossa, a do PHPB/BA. Os alvos são três: *corpus, história social e análise de mudança lingüística*. **A formação do *corpus* pressupõe edição conservadora de documentos fidedignos**, **localização temporal**, **geográfica e o perfil biográfico dos autores dos documentos**, se brasileiros, portugueses, data de nascimento, uma vez que se acredita que a mudança se dá no processo de aquisição. Para esse estudo faz-se fundamental a contribuição da micro-história, como a desenvolvida pelo historiador Erivaldo Fagundes Neves, pesquisador da UEFS e colaborador do projeto da Equipe UEFS.

2. Da formação e consolidação da mudança lingüística

A concepção da gramática gerativa, construída segundo o pressuposto de que há princípios universais e princípios parametrizáveis responsáveis pela variação que se observa de língua para língua, permitiu que a mudança adquirisse um novo enfoque dentro dessa teoria. A conclusão imediata dessa formulação é que a mudança se daria durante o processo de aquisição da linguagem. Essa formulação teve grande aceitação e logo se tornou consensual. O problema passou a ser o de dimensionar a forma como a experiência lingüística atua nesse processo. Através do modelo de Princípios e Parâmetros (Chomsky, 1981), construído com base da comparação das línguas, é possível interpretar os fatos lingüísticos encontrados em textos diacrônicos e lhes atribuir uma gramática subjacente. A mudança sintática é interpretada como a mudança na fixação paramétrica.

Assim sendo, a relação entre mudança paramétrica e aquisição da linguagem motivou diversas tentativas de elaboração de uma teoria de mudança que viesse a dar uma resposta adequada a um dos problemas cruciais dessa teoria: o que leva uma criança a marcar diferentemente dos seus pais, ou da geração anterior, os parâmetros da língua que lhe serviram de *input*. Por outro lado, o fato de a lingüística histórica não ter acesso a dados introspectivos, mas apenas dados de língua-E foi alvo de discussões importantes na linha de interpretação da gramática gerativa. Para os pesquisadores dessa área, a mudança paramétrica é por definição abrupta diferindo de outras interpretações tradicionais que defendem a sua natureza lenta e gradual, expressa através de processos de coexistência e concorrência das formas variantes pelo período que a antecede.

Para resolver esse impasse, estudiosos na área da gerativa defendem que a mudança é gradual nos textos porque a mudança paramétrica é o fim de um processo, como Roberts (1993) que reinterpreta as fases da lingüística histórica tradicional e diz

que é possível ver isso através de alterações de freqüência, baixa na freqüência, reanálise e o seu desaparecimento. Ou seja, em seus termos, a mudança paramétrica é captada nessa última fase. Do seu ponto de vista, assim como o de Lightfoot (1999), nesse momento é que a mudança se torna abrupta, daí ser vista como o fim de um processo. Por outro lado, Kroch (1994, 2001), ao associar a relação entre língua-I e alterações de freqüência, defende que o que se vê nos textos é a tensão entre a gramática nova e a gramática antiga captada através de alterações na freqüência que atingem um conjunto de propriedades associadas a um determinado parâmetro, resultando na denomina Hipótese da Taxa Constante. Segundo essa assunção o que aparece nas amostras de língua-E é apenas o efeito da mudança.

A distinção entre língua-I e língua-E, fundamental para o programa de investigação da gramática gerativa, que tem como objeto de estudo a língua-I, ganha especial relevância nos estudos diacrônicos, uma vez que as mudanças definidas a partir dessa concepção de gramática são tratadas como alterações paramétricas.

Essas alterações seriam decorrentes de falhas de transmissão lingüística durante o processo de aquisição da linguagem por crianças (língua materna ou L1), ou por adultos em situação de contato lingüístico (segunda língua ou L2).

3. A análise de fenômenos lingüísticos: estudos diacrônicos sobre o português brasileiro

Além dessas questões colocadas acima, há um aspecto importante que marca os estudos diacrônicos no Brasil no âmbito da gramática gerativa, o de interpretar dados quantificados em amostras de língua-E com base nas hipóteses sobre gramáticas abstratas apreensíveis nos textos. Os resultados desses estudos trouxeram um avanço teórico significativo para a compreensão do português brasileiro. Uma das mais importantes contribuições deve-se à hipótese sobre mudanças no sistema pronominal brasileiro. Essas mudanças estariam relacionadas à reestruturação do sistema de caso e ao surgimento de uma gramática distinta do português entre o século XVII e o século XIX.

Esses projetos lingüísticos pautam-se em importantes linhas de investigação da lingüística atual que buscam compreender, com dito acima, o que provoca a mudança lingüística, e como essa mudança se dá ao longo do tempo. Como contraparte social final, ajudar os educadores a entender a língua com a qual trabalham metalingüisticamente nas salas de aula, quando o resultado desses estudos chega aos livros didáticos através de projetos pedagógicos.

Os estudos em lingüística histórica, numa perspectiva gerativa, utilizam os dados reais de língua-E para extrair uma gramática abstrata, a língua-I. A metodologia impõe **o como analisar as construções de língua-E**. Por exemplo, se ao filólogo, no cumprimento de suas funções, cabe a preocupação com a autenticidade do texto, localização espacial e temporal, etc., ao lingüista cabe saber se se trata de documentos escritos por pessoas que têm essa língua como materna (L1) ou segunda língua (L2), uma vez que se defende que a gramática é construída durante a aquisição da linguagem. Por outro lado, o conhecimento prévio do que determina uma mudança paramétrica é também crucial para que as construções que a representam, normalmente pouco representativas nos textos, não sejam ignoradas, principalmente em se tratando de amostras de língua escrita, por natureza conservadora e que podem estar refletindo a norma ou a

tradição escrita de um dado período. A separação entre usos provocados por influências artificiais como, por exemplo, as influências de cunho estilístico ou discursivo, é importante para a identificação da mudança relevante em análises a partir de textos escritos, como é o caso desta pesquisa.

Atualmente no Brasil, há já muitos resultados com de estudos de diversas linhas teóricas e mostrando as propriedades específicas do português brasileiro, atestadas por diversos pesquisadores em resultados quantitativos e baseados em *corpora* diacrônicos da segunda metade do século XVII-XX (cf. Galves, 1987; Tarallo, 1989 e Roberts e Kato; 1993). Daí a vantagem do corpus digital eletrônico por permitir que se trabalhe com grande volume de dados para quantificação.

4. Da formação de bancos eletrônicos digitais

Para atingir uma análise qualitativa através da análise quantitativa é que tem se buscado aliar o recurso a grandes volumes de dados, através da contribuição de tecnologias computacionais, como a desenvolvida pelo "Projeto Corpus do Português Histórico Tycho Brahe" e do projeto interdisciplinar "Rítmicos, Fixação de Parâmetros e Mudança Lingüística - Fase II", a partir de investigações teóricas solidamente fundadas, "reunindo e fomentando a contribuição entre pesquisadores de formação variada - sintaticistas, fonólogos, estatísticos, probabilistas, teóricos da computação e lingüistas computacionais - contribuindo, por fim, na renovação das perspectivas para o campo", como especificado nessa fase do projeto no seguinte endereço http://www.tycho.iel.unicamp.br/~tycho/prfpml/fase2/projeto_completo.html). Essa vem se mostrando uma tendência mundial, como podemos ver a partir do número de projetos elencado a seguir: o pioneiro *Penn Helsinki Parsed Corpus of Middle English* , (http://www.ling.upenn.edu/hist-corpora), coordenado por Anthony Kroch na Universidade da Pensilvânia e os seus afiliados o *York Helsinki Parsed Corpus of Old English Poetry*, por Susan Pintzuk e Leendert Plug, o *York Toronto Helsinki Parsed Corpus of Old English Prose,* por Ann Taylor, Anthony Warner, Susan Pintzuk, Frank Beths, ambos na Universidade de York, o *Penn Helsinki Parsed Corpus of Early Middle English,* por Anthony Kroch e Beatriz Santorini na Universidade da Pensilvânia, e enfim o *Parsed Corpus of Early English Correspondence,* por Ann Taylor, Anthony Warner, Susan Pintzuk na Universidade de York, e por Terttu Nevalainen e Arja Nurmi na Universidade de Helsinki, além de outros como o projeto de *Corpus annoté syntaxiquement de textes de français (9è au 17è siècle),* por F. Martineau e Paul Hirschbuhler na Universidade de Ottawa, e o projeto *Corpus Dialetal Sintático (CordialSin),* por Ana Maria Martins, na Universidade de Lisboa. Diante desse quadro vê-se que os avanços tecnológicos são iminentes. De modo geral, tem se usado a linguagem XML pelas vantagens que apresenta (cf. apresentado metodologia)

Para finalizar, é importante salientar que esse tipo de edição eletrônica em linguagem XML e seus produtos abrem uma grande janela de oportunidades para edições filológicas com a aplicação eletrônica de textos, por possibilitar diferentes "camadas de informação em um texto" ao invés do uso não manipulável da digitação em Word (Paixão de Sousa, 2006). O uso dessa tecnologia poderá dar novo impulso a Lingüística Histórica.

METAS

1) Elaboração do plano central do Piloto de Corpus Digital Eletrônico.
2) Aprender a explorar o EDICTOR.
3. Codificação em XML 185 textos da estrutura bruta em Word Piloto de Corpus Digital Eletrônico.
4) Elaboração do modelo para os módulos automáticos de busca no Piloto de Corpus Digital Eletrônico.
5) Revisão final da estrutura para teste de diretórios do Piloto de Corpus Digital Eletrônico.

6. Elaboração de artigo de análise lingüística como resultado da aplicação do Piloto de Corpus Digital Eletrônico.
6. Redigir relatórios finais

PLANO DE ATIVIDADES:
Semestre 1:
- Estudo detalhado das possibilidades de elaboração e aplicação do plano geral do "Corpus Histórico do Português Tycho Brahe" ao Piloto de Corpus Digital Eletrônico.
Aprofundamento da pesquisa de técnicas de anotação XML; estudo detalhado das diretrizes do Piloto do Corpus Digital; esboço das diretrizes de anotação.
- Anotação da estrutura bruta e metadata parcial 50 textos;-
Participação em **oficinas** com as demais equipes do projeto;
Apresentando o esboço para as diretrizes de preparação e a anotação XML dos primeiros 50 textos;
Semestre 2:
Preparação de uma primeira versão do Piloto do Corpus Digital Eletrônico
extensão da anotação para todos os textos;
Re-estruturação da arquitetura do corpus eletrônico (a partir da página web) do Piloto do Corpus Digital Eletrônico
- Preparação de uma oficina a ser aplicada na UEFS, após o meu retorno para treinamento de equipes para alimentação do Corpus Digital a partir do piloto.
- Elaboração do arcabouço geral do site na UEFS do Piloto do Corpus Digital

Referências Bibliográficas:

BARBOSA, Afrânio Gonçalves / LOPES, Célia (Orgs. 2002 / 2006). Críticas, Queixumes e Bajulações na Imprensa Brasileira do séc. XIX: cartas de leitores e cartas de redatores. Rio de Janeiro: Projeto para a História do Português Brasileiro / Universidade Federal do Rio de Janeiro, cd-rom. Versão em papel: Rio de Janeiro: Universidade Federal do Rio de Janeiro / Faperj, 2006.

BRITTO, Helena & FINGER, Marcelo (1999): *Constructing a parsed corpus of historical portuguese.* ACHALLC' 99 International Humanities Computing Conference, University of Virginia, Charlottesville, 9-14 de junho de 1999.

CARNEIRO, Z. O. N. (2008). Vozes do sertão em dados: história, povos e formação do português brasileiro. In: *VI Feira do Semi-Árido: desertificação, perspectivas de autonomia produtiva frente aos desafios socioambientais,* 2008, Feira de Santana. VI

Feira do Semi-Árido: desertificação, perspectivas de autonomia produtiva frente aos desafios socioambientais. Feira de Santana, v. 1.

CARNEIRO, Z. O. N. (2008). Estudo de escolarização de aldeados no Brasil do século XVII: um caminho para a compilação de possíveis fontes escrita em português por. In: XV Congresso Internacional de La Asociación de Lingüística y Filología de América Latina, 2008, Montevideo. Libro de resúmenes de XV Congresso Internacional de La Asociación de Lingüística y Filología de América Latina/ALFAL, p. 263-263.

CARNEIRO, Zenaide Novais (2005). *Cartas brasileiras (1809-1907): um estudo filológico-linguístico.* Campinas: UNICAMP. Tese de doutorado inédita.

CARNEIRO, Zenaide de Oliveira Novais; ALMEIDA, Norma Lucia F. de (2006). A criação de escolas a partir de critérios demográficos na Bahia do século XIX: uma viagem ao interior. In: LOBO, Tânia; RIBEIRO, Ilza; CARNEIRO, Zenaide de O. N.; ALMEIDA, Norma Lucia F. de. *Para a história do português brasileiro: novos dados, novas análises*. Salvador: Edufba, vol. 6, 1-2, p. 649-673.

CARNEIRO, Zenaide de Oliveira Novais; ALMEIDA, Norma Lucia F. de. (2007. Elementos para uma sócio-história do semi-árido baiano. In: RAMOS, J.; ALKMIM, Mônica A. *Para a história do português brasileiro*: *estudos sobre mudança lingüística e história social*. Belo Horizonte: Faculdade de Letras da UFMG, v.5. p. 423-442.

CASTILHO, Ataliba T. (Org., 1998). Para a História do Português Brasileiro, vol. I, Primeiras Idéias. São Paulo: Humanitas /Fapesp.

CES, 1996: Corpus Encoding Standard, Document CES 1 Version 1.2 . Em <http://www.cs.vassar.edu/CES/CES1-1.html>
CHOMSKY, Noan. (1986a). *Knowledge of language:* Its nature, origin and use. New York: Praeger.

CHOMSKY, Noan. (1988). *Language and problems of knowledge*. The Managua Lectures. Cambridge: MIT Press.

CHOMSKY, Noan. (1995). *A minimalist program*. Press. Massachusetts: Cambridge, MIT PRESS.

FINGER, Marcelo, 1998: "Tagging a morphologically rich language", *Proceedings of the first workshop on text, speech and dialogue* (TSD98), Brno, Tchecoslováquia.

FINGER, Marcelo, 2000: *Técnicas de otimização da precisão empregadas no etiquetador Tycho Brahe*,
Propor 2000.

GALVES, C. (1987). A sintaxe do português brasileiro. *Ensaios de lingüística,* p. 13:31-50.

GALVES, C. (2002). *Ensaios sobre gramáticas do português.* Campinas: Editora da UNICAMP.

GALVES, Charlotte, 2004: *Projeto Padrões Rítmicos, Fixação de Parâmetros e Mudança Gramatical, II* –FAPESP.
GALVES, Charlotte

(http://www.tycho.iel.unicamp.br/~tycho/prfpml/fase2/index.html) relatórios anuais.

GUEDES, Marymarcia & BERLINCK, Rosane Andrade (Orgs. 2000). E os preços eram commodos…Anúncios de jornais brasileiros do século XIX. São Paulo: Humanitas [Série Diachronica, vol. 2].

IDE, Nancy e Laurent Romary, 2003: *Outline of the international standard linguistic annotation*
framework. Proceedings of ACL'03 Workshop on Linguistic Annotation: Getting the Model Right, Sapporo, 1-5. Em <http://www.cs.vassar.edu/faculty/ide/pubs.html>

ROBERTS, I; KATO, M. (Orgs). *Português brasileiro*: uma viagem diacrônica. Campinas: Editora da UNICAMP,

KROCH, A. (1989b). Reflexes of grammar in patterns of language change. *Language variation and change,* 1, p. 199-244.

LABOV, W. (1982). Building on empirical foudation. In: LEHAMANN, W. P.; MAKIEL, Y (eds.). *Perspectives on historical linguistics*. Philadelphia: John Benjamins Publishing Company.

LEITE, S.J., Serafim (1938-1950). *História da Companhia de Jesus no Brasil*, 10 vols. Lisboa e Rio de Janeiro: Portugália/Civilização Brasileira.

LIGHTFOOT, David. (1999). *The development of language:* Acquisition, change, and evolution. Maryland lectures in language and cognition. Malden.Blackwell.

LOBO, Tânia / RIBEIRO, Ilza Ribeiro / CARNEIRO, Zenaide / ALMEIDA, Norma (Orgs. 2006). Para a história do português brasileiro: novos dados, novas análises. Salvador: Editora da Universidade Federal da Bahia, vol. VI, 2 tomos.

LOBO, Tânia; MATTOS E SILVA, Rosa Virgínia; VENÂNCIO, Américo L. M. Filho (2006). Indícios de uma língua geral no sul da Bahia na segunda metade do século XVIII. In: LOBO, Tânia; RIBEIRO, Ilza; CARNEIRO, Zenaide de O. N.; ALMEIDA, Norma Lucia F. de. *Para a história do português brasileiro*: *novos dados, novas análises*. Salvador: Edufba, vol. 6, 1-2, p. 609-630.

LOBO, Tânia (2003). A questão da periodização da história do Brasil. In: CASTRO, Ivo; DUARTE, Inês (orgs.). *Razão e emoção: miscelânea de estudos em homenagem a Maria Helena Mateus*. Lisboa: Imprensa Nacional- Casa da Moeda, p.395 et.passim.

MARQUILHAS, Rita, (2001). *A faculdade das letras*, Imprensa Nacional Casa da Moeda, Lisboa.
MATTOS E SILVA, Rosa Virgínia (2008). *Caminhos da Lingüística histórica: ouvir o inaudível,* São Paulo: Parábola Editorial.

MATTOS E SILVA, Rosa Virgínia (2001). De fontes sócio-históricas para a história social lingüística do Brasil: em busca de indícios. In: MATTOS E SILVA, Rosa Virgínia (org.). *Para a história do português brasileiro: primeiros estudos*. São Paulo: Humanitas/FFCHL/USP:FAPESP, v.2, t. 2, p. 275-302.

MATTOS E SILVA, Rosa Virgínia; OLIVEIRA, Klebson; LOBO, Tânia (2007). Panorama preliminar do letramento de negros na Bahia. In: RAMOS, J.; ALKMIM, Mônica A. *Para a história do português brasileiro*: *estudos sobre mudança lingüística e história social*. Belo Horizonte: Faculdade de Letras da UFMG, v.5. p. 373-442.

MATTOS E SILVA, Rosa Virgínia (2002). Para a história do português culto e popular brasileiro: sugestões para uma pauta de pesquisa. In: ALKMIM, Tânia M. *Para a história do português brasileiro*: *novos estudos*. São Paulo: Humanitas/FFCHL/USP:FAPESP, v. 2, p. 443-464.

MATTOS E SILVA, Rosa Virgínia (Org.2001). Para a História do Português Brasileiro, vol. II, Primeiros Estudos, 2 tomos. São Paulo: Humanitas / Fapesp.

MONTEIRO, John M. (2001). *Tupis, tapuias e historiadores*: *estudos de história indígena e do indigenismo*. Campinas: IFCH/UNICAMP. Tese de livre docência.

PAIXÃO DE SOUSA, M.C., CAVALCANTE, S.R.O., NAMIUTI, C.. *Lingüística de Corpus e História da Língua Portuguesa: Propostas, Resultados e Desafios.* Mesa Redonda. V Congresso Internacional da ABRALIN. 2007.*História da Língua Portuguesa: Propostas, Resultados e Desafios.* Mesa Redonda. V Congresso Internacional da ABRALIN. 2007.

PAIXÃO DE SOUSA, M.C. *Memórias do Texto*. Revista Texto Digital. n. 2. Universidade Fedral de Santa Catarina. 2006.
PAIXÃO DE SOUSA, Maria Clara. *Projeto Memórias do Texto*. FAPESP-UNICAMP, 2004.

PAIXÃO DE SOUSA, Maria Clara e Thorsten Trippel (2004): *Single source processing of historic corpora for diverse uses* ALLC/ACH 2004, proceeding.

PROJETO: Os Índios na História do Brasil: Informações • Estudos •Imagens, coordenador por J. M. Monteiro/IFCH/UNICAMP/http://www.ifch.unicamp.br/ihb.

RAMOS, Jânia / ALCKMIN, Mônica A. (Orgs. 2007). Para a História do Português Brasileiro, vol. V: Estudos sobre mudança linguística e história social. Belo Horizonte: Faculdade de Letras da Universidade Federal de Minas Gerais.

SAXON. <http://saxon.sourceforge.net/>
TYCHO BRAHE. <http://www.tycho.iel.unicamp.br/~tycho/corpus/index.html>
TRIPPEL, Thorsten and PAIXÃO DE SOUSA, Maria Clara (2006). "Metadata and XML standards at work: a corpus repository of Historical Portuguese texts". *Papers from the V International Conference on Language Resources and Evaluation (LREC 2006).*

W3C (1997). "Extensible Markup Language". http://www.w3.org/XML